PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

23. října 2012

WSD, pokračování

Word Sense Desambiguation

úkolem WSD je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

minule jsme mluvili o metodách založených na znalostech (Leskův algoritmus pracující se slovníkovými definicemi a příklady užití)

Algoritmy strojového učení

- „matematické“
  - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
  - maximální entropie: Berger 1996
  - podobnostní: vector space model, k-NN (Ng et Lee, 1996)
- „promluvové“
  - one sense per discourse (Gale 1992)
  - one sense per collocation (Yarowsky, 1995)
  - redundance atributů
- „pravidlové“
  - rozhodovací seznamy
  - rozhodovací stromy

Nebudeme se tu nějak moc věnovat ML, ale přeci jen poskytnu povrchní přehled. S uvedenými termíny se totiž může počítačový lingvista poměrně často setkat, tak ať aspoň ví, na čem je.

Pro lidi jsou typické (a intuitivní) spíš „pravidlové“ systémy.

Příkladem je hra Myslím si zvíře (protihráč se snaží zvíře $z \in \mathcal{Z}$ uhádnout pomocí otázek, na které dostává odpovědi ano/ne).

V této hře jsou 2 aspekty:

- Jak poznám z množiny otázek $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$, kde o_i je např. „Má zvíře srst?“, o jaké zvíře jde? Redukcí. Pokud odpověď na o_i je „ne“, vyloučím ze správných odpovědí všechna zvířata z_j, která mají srst. Podobně pro další otázky, dokud nezůstane (ideálně) 1 zvíře.
- Jaká je strategie kladení otázek? Cílem je minimalizovat n. Prostředkem k dosažení tohoto cíle je neklást otázky, které dělí množinu možných zvířat stejným způsobem. Např. otázky „Má zvíře srst?“ a „Má zvíře 4 nohy?“ dělí $\mathcal{Z}$ na dvě téměř stejné části.

Na celou hru můžeme pohlížet jako na množinu zvířat (které známe) a rozhodovací strom, který nás „dovede“ k vítěznému zvířeti.

Algoritmus strojového učení [Yarowsky, 1995]

hledáme význam *s* slova *w*

- 1. vezmi všechny výskyty slova *w* z korpusu včetně jejich kontextů
- 2. pro každý možný význam slova, vytvoř malou sadu příkladů (buď ručně, nebo pomocí kolokací)
- 3. vytvoř rozhodovací seznam s pravděpodobnostmi pro další slova, která se vyskytují v kontextech
- 4. aplikuj tento seznam na celý korpus (s prahem pro pravděpodobnost)
- 5. nově zařazená slova obsahují další slova v kontextech
- 6. algoritmus můžeme upravit pomocí zařazení předpokladu one-sense-per-discourse
- 7. opakuj kroky 3–6
- 8. jakmile množiny přestanou narůstat, zastav
- 9. systém je nyní natrénovaný i na jiný korpus!

Algoritmus strojového učení

závisí na:

- první volbě kolokací
- způsobu určení pravděpodobnosti: typicky log likelihood $\log \frac{P(senseA,collocateA)}{P(senseB,collocateA)}$
- prahu pro pravděpodobnost
- správnosti předpokladu one-sense-per-discourse


Yarowsky, D. (1995).
Unsupervised word sense disambiguation rivaling supervised methods.
In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.